REPLICATING WEB STRUCTURE IN SMALL-SCALE TEST COLLECTIONS

CATHAL GURRIN cgurrin@computing.dcu.ie
ALAN F. SMEATON asmeaton@computing.dcu.ie
Centre for Digital Video Processing, Dublin City University, Glasnevin, Dublin 9, IRELAND

Abstract: Linkage analysis as an aid to web search has been assumed to be of significant benefit and we know that it is being implemented by many major Search Engines. Why then have few TREC participants been able to scientifically prove the benefits of linkage analysis in recent years? In this paper we put forward reasons why many disappointing results have been found in TREC experiments and we identify the linkage density requirements of a dataset to faithfully support experiments into linkage-based retrieval by examining the linkage structure of the WWW. Based on these requirements we report on methodologies for synthesising such a test collection.

Key words: linkage analysis, search engine, retrieval evaluation, test collections

1. INTRODUCTION

The first generation of web search engines which have contributed to the huge popularity of the WWW were based on directly computing the similarity between a query and the text appearing in a web page and were effectively a direct application of standard document retrieval techniques. While these initial "first generation" web search engines addressed the engineering problems of web spidering and efficient searching for large numbers of both users and documents, they did not innovate much in the approaches taken to document ranking.

In the past few years we have seen most, if not all, web search engines incorporate linkage analysis as part of their retrieval operation. Anecdotally

this appears to have improved the precision of retrieval yet, up until recently, there has been little scientific evidence in support of the claims for better quality retrieval, especially using the conventional TREC evaluation methodology of using test collections to evaluate retrieval performance. Test collections are an essential aspect of retrieval performance evaluation for web search tasks because firstly, it is very difficult for numerous different research groups to carry out comparable and repeatable experiments on live web data and secondly, test collections are generally constructed to be small in size (1-2 million documents) so as to encourage their widespread use in IR experimentation.

Participants in the four most recent TREC conferences[1] (1999 - 2002) have been invited to perform benchmarking of information retrieval systems on web data and have had the option of using linkage information as part of their retrieval strategies. Until TREC-2002, the general consensus was that except in extremely rare cases, and for insignificant improvements anyway, linkage information had not yet been successfully incorporated into conventional retrieval strategies when evaluated using the TREC test collection methodology. In most cases, linkage-based information was found to significantly harm conventional retrieval though improvements had been found specifically in a homepage finding task, the aim of which was to locate homepages contained within the two TREC test collections used in 2001 and 2002.

In this paper we present a rationale as to why we believe TREC Web Track participants (including ourselves at DCU) using the TREC web-based test collections (prior to 2002) have been unable to demonstrate improved retrieval performance when incorporating linkage analysis into retrieval strategies and why this was not the case for some TREC participants in 2002 who were able to demonstrate improvements in retrieval performance. We begin with a brief introduction to linkage analysis in the next section and follow that with an overview of the TREC web track focusing on an analysis of the datasets employed, including a brief description of our own experiments on a test collection extracted from one of the TREC collections. We then report on the results of an examination of the linkage structure of the crawlable WWW (i.e. the section of the WWW that can be indexed by search engines) and based on these findings we suggest the linkage requirements for any future datasets to support TREC style evaluation of linkage-based retrieval. We finish up by reporting methodologies for synthesising such a test collection that meets our suggested linkage requirements and our progress at actually generating such a collection.

[1] At the time of writing TREC-2003 has not yet taken place and no results are available yet, although we note that web retrieval is once again under experimentation in 2003.

2. AN INTRODUCTION TO LINKAGE-BASED RETRIEVAL

The "first generation" of search engine primarily utilised the content of the document when generating ranked listings. However, an additional source of latent information available to web collections is how documents are linked together and it is the study of this aspect of the web that is referred to as linkage analysis. More recent search engines[2] utilise linkage data and are able to gather information mined from the documents themselves, as well as information from the linkage structure of the web. In most cases this linkage information is represented as a 'connectivity' or a 'linkage' score for each document, which will influence final document ranking.

Generally speaking, on the WWW we can separate links into one of two broad types based on their intended function when created:

- *On-site* links are created to link documents within a particular domain and exist to aid the user in navigating within a domain, or website. Thus they may be described as being either purely navigational or informational in nature. While information can be mined from these on-site links, which may aid certain aspects of retrieval, this is not often done because of the difficulties in separating informational from navigational links.
- *Off-site* (content, or outward) links on the other hand link documents from different domains (across web site boundaries). They are found to mostly link from a source document to a target document that contains similar and, in the web page author's opinion, useful information. Many linkage-analysis techniques focus on off-site links at the expense of on-site links.

In general we can assume that a document with a higher number of off-site in-links (where in-links are hypertext links pointing to a web page) will be a more 'popular' document than one with less off-site in-links. For the purpose of linkage analysis we are interested primarily in the number of (and the quality of) off-site citations (in-links) that a web page receives. One reason for this is that many on-site links are of the purely navigational type and it would be difficult to reliably separate these navigational on-site links from the more useful informational on-site links. If a web page receives a large number of off-site links then we can broadly conclude that this page may be a better page than one that receives significantly fewer off-site links. Thus, in the context of linkage information for conventional web searching, we are primarily interested in off-site links as opposed to on-site links.

[2] One possible exception here is the WWWW (McBryan, 1994), possibly the first true WWW search engine which utilised anchor text in the retrieval process.

Given that an off-site link to a document can be seen as an indication of the usefulness of that document, a simple linkage score can be generated for each document in a set of highly-scored documents based on the off-site indegree of each document and hence we can rank documents by off-site indegrees. Researchers at AT&T (Amento et al. 2000) have demonstrated, using their own crawled data, that incorporating indegree ranking into retrieval strategies is equally as effective as incorporating other more advanced techniques for using linkage information, such as the PageRank algorithm.

PageRank (Page et al. 1997) is probably the best-known linkage analysis technique in use on the web today and is believed to be implemented in the Google search engine (Brin and Page 1998). PageRank is based on a simple indexing-time process that generates a linkage score (the PageRank) for each document in the search engine's index. This PageRank score is combined at query-time with other sources of evidence such as a content-only score to produce a final document score used in ranking results. PageRank is based on a simulation of a random user's behaviour while browsing the web where the user keeps clicking on successive links at random. Due to the fact that a user can get caught in page loops, rather than looping forever the user jumps to a random web page (chosen using the vector E over all web pages). E is normally uniform for all web pages.

PageRank is calculated over a number of iterations until an acceptable level of convergence of the PageRanks has been reached. The PageRank (*Pr'*) of a document is calculated using a formula similar to the following (see formula 1), where S_n is the set of documents that link into document n, c is a constant used for normalisation, Pr_n is the current PageRank of n and E is a uniform vector over all web pages.

$$\mathrm{Pr}'_n = c \cdot \sum_{m \in S_n} \frac{\mathrm{Pr}_m}{\mathrm{outdegree}_m} + (1-c) \cdot E_n \qquad (1)$$

Another well-known technique incorporating linkage information in web searching is Kleinberg's (Kleinberg 1999), which is similar to PageRank in that it is an iterative algorithm based purely on the linkage between documents but it has major differences, namely:

- It is executed at query time, rather than at indexing time.
- It computes two scores per document (Hub and Authority) as opposed to one single score. Hub scores reflect a document's usefulness as a source of links to relevant content while Authority scores represent a document's usefulness as a source of relevant content.
- It is processed on a small subset of 'relevant' documents, not all documents.

The Hub and Authority scores are calculated for each document on a small subset chosen due to their rank in a content-only search run, or due to their being in the immediate neighbourhood of these highly ranked documents. The process is iterative and the Authority and Hub vectors will eventually converge, at which point the iterations can stop. Once convergence has been achieved, the documents are ranked into two groups, by Hub (links to content) and Authority (content) scores. A number of improvements to this model have been suggested and successfully evaluated (Bharat and Henzinger 1998).

3. EVALUATING LINKAGE-BASED RETRIEVAL: THE TREC WEB TRACK

By 1999, linkage analysis and web search in general had become a 'hot' topic and a special 'web track' in the annual TREC benchmarking exercise was dedicated to web search related tasks. From 1999 to 2001 this web track has supported TREC participants in their endeavours to find out whether the best methods in ad-hoc (conventional) retrieval also work best on the TREC collections of web data and whether link information in web data can be used to obtain more effective retrieval than using page content alone (Hawking et al. 1999).

In 2002, this conventional, or ad-hoc search task of the TREC web track was replaced by a Topic Distillation task (evaluated using the measure of precision at 10), the goal of which is to find a small number of key resources on a topic as opposed to the more conventional (ad-hoc) listing of relevant pages. Topic Distillation, although not that far removed from ad-hoc is perhaps more suited to web search evaluation because it has been found that over 85% of users never look beyond the first page of results from any web search (Silverstein et al. 1998).

In order to support the experiments of participants, TREC distributes test collections that consist of three components, a set of documents, a set of queries (called topics) and a set of relevance judgements for each query. The first TREC test collection for the web track was the WT2g collection that was used in the TREC-8 conference in 1999. This is a 2GB collection and was said to contain an "interesting quantity" of closed hyperlinks (having both source and target within the dataset). A larger collection of 10 GB of web data, known as WT10g, was used in TREC-9 and TREC-2001. Most recently an 18 GB collection (.GOV) was used for TREC 2002. We now examine each collection.

3.1 WT2g (TREC-8)

As stated above, the TREC-8 web track used a 2GB collection called WT2g[3] that consisted of 247,491 HTML documents. The WT2g collection is a subset of the 100GB VLC dataset[4], which is itself a subset of a 300 GB Internet Archive crawl completed in early 1997. Seventeen participating groups took part in the web track in TREC-8 and those that utilised the link information implemented a variety of approaches including Kleinberg's and PageRank mentioned earlier. We ourselves implemented techniques based on citation counting (Gurrin and Smeaton 1999).

To the initial surprise of many participants, no participating group (save one with insignificant improvements) managed to improve precision over that attained by conventional content-only searching when linkage information was incorporated into the retrieval process. There were a number of reasons put forward to explain this (Hawking 2000) but the primary reason seemed to be the sparcity of linkage data within WT2g. The number of closed off-site links within WT2g is 2,797 out of 1,166,702, which effectively means that there is a maximum of one off-site link for every 100 documents. A consequence of this is that we, and other participants, found WT2g to be incapable of demonstrating any benefits by incorporating linkage algorithms into the retrieval process.

3.2 TREC-9 & TREC-2001 (WT10g)

In a manner similar to TREC-8, an ad-hoc retrieval task was employed for TREC-9 and TREC-2001. The shortcomings of WT2g had led to the creation of a new collection, WT10g, which was used in the web tracks of TREC-9 and in TREC-2001. A collection size of 10GB was chosen which comprised 1,692,096 documents. Similar to the preceding WT2g, WT10g was also subset of the 100GB VLC dataset but was extracted from the VLC in such a way that it met certain pre-defined criteria (Bailey et al. 2003) primarily based on replicating server size distribution figures, but also keeping the number of off-site links included within WT10g high. However, the linkage density of the VLC itself served to restrict the number of links that are candidates for inclusion in WT10g. Time constraints (Bailey et al. 2003) had precluded the use of the whole 320GB Internet archive crawl and therefore, the 100GB VLC collection was employed instead.

[3] WT2g means Web Track 2 gigabytes.

[4] VLC, Very Large Collection, which was used in the large web task from 1999 to 2001. Note that in this paper, we focus on the small web task, not the large web task.

The WT10g collection contained a total of 171,740 off-site links for 1.69 million documents, averaging at one off site link for every ten documents. Thus any linkage-based techniques could only influence a small proportion of the documents within WT10g and once again none of the participants' experiments (22 participants in all) for TREC-9, including our own experiments on citation analysis and spreading activation (Gurrin and Smeaton 2000), were able to illustrate any improvement in performance over content-only retrieval when incorporating any linkage-based algorithm into the retrieval process. For a more detailed description of WT10g, we refer the reader to a description of how WT10g was constructed (Bailey et al. 2003).

TREC-2001 (the 10th TREC conference) once again encouraged participants (29 in all) to partake in linkage analysis experiments using WT10g and once again linkage analysis was not found to aid retrieval performance in the ad-hoc retrieval task. This time, in addition to the conventional retrieval task, a new task was introduced to the web-track, namely a homepage finding task (essentially a known-item task) which it was hoped would better support linkage-based methods than would the conventional retrieval task. Prior evaluation of such methods during construction phase of WT10g had illustrated that such a task could indeed yield performance improvements (Bailey et al. 2003) and some linkage-based improvements were indeed evident in this task, although not in the ad-hoc retrieval task, in which we are most interested.

3.3 WT-Dense: A Densely Linked Subset of WT10g

We feel that the WT10g dataset seriously underestimates the density of off-site links, which does not allow it to support faithful experiments on linkage-based web searching. However, the primary advantage of using WT10g was readily available relevance judgements for 100 queries. Generating a new dataset would require undertaking the relevance judgement process, which is both expensive and time-consuming. Our solution was to develop a subset of WT10g (including relevance judgements) in order to evaluate our linkage algorithms. We used the TREC-9 query set for these experiments (topics 450-499).

When generating this densely linked subset of WT10g (called WT-Dense) we had two requirements for the new collection, these being to maximise the number of off-site links in the dataset, and to maximise the size of the new dataset itself. Generating a dataset to satisfy these two requirements was straightforward and is detailed elsewhere (Gurrin and Smeaton 2003). A comparison of the linkage structure between WT-Dense and WT10g is now summarised:

Table 1. Comparing WT10g and WT-Dense

	WT10g	WT-Dense
Number of Documents	1,692,096	120,494
Number of off-site links	171,740	171,740
Average off-site indegree	0.10	1.43
Number of unique servers represented	11,680	11,611
Generality	0.15%	0.21%
Number of TREC-9 Queries with relevant documents	50	36
Average number of relevant documents per query	52	7

As can be seen from Table 1, WT-Dense contains a far higher density of off-site links, an average of 1.43 per document while keeping the generality (percentage of documents relevant to any of the topics) of the dataset similar to WT10g. However, results of a survey of web structure presented later in this paper suggest that this average off-site indegree figure of 1.43 is actually less than one third of the true off-site indegree figures as found on the WWW.

Fourteen of the fifty TREC-9 queries had no relevant documents in WT-Dense, which reduced the number of usable queries to 36, thus reducing the number of performance comparisons. In addition, although unavoidable, the average number of relevant documents per query was reduced to 7, which we note to be well below the norm.

3.3.1 Experiments on WT-Dense

We ran a number of retrieval experiments (content-only and linkage based) using WT-Dense for the TREC-9 topics using manually generated queries. Our first experiment was a content-only experiment for which we utilised the popular BM25 ranking algorithm. The top 1,000 ranked documents for each query were used as a benchmark against which to compare the retrieval performance of our subsequent linkage-based runs. These subsequent linkage-based runs were based on re-ranking the top 2,000 documents produced in the content-only phase using linkage analysis algorithms based on:

- *Citation ranking*, off-site indegree ranking using normalised indegree scores with two methods of combining linkage and content evidence. Firstly using parameter combination (a linkage weight of 0.25 and a content weight of 1.0) and secondly using our own technique, which we call the scarcity-abundance technique. Essentially scarcity-abundance dynamically estimates linkage and content influence using a broadness measure (based on the size of the content-result set) for each query. For more details see (Gurrin and Smeaton 2003). This experiment is the

results section (3.3.2) is referred to as 'Citation A' (parameter combination) and 'Citation B' (scarcity-abundance) combination.

- *Spreading activation*, a technique that propagates numerical values (or activation levels) among the connected nodes of a graph. In the context of this experiment it facilitates a document transferring its content-only score across its out-links. Only documents that are scored in the content-only phase have any effect on the process and this is referred to as SpreadAct in the results section.
- *PageRank*, we also evaluated the PageRank algorithm. We had calculated a single PageRank score for each document in WT-Dense, prior to running these experiments, using the PageRank algorithm as outlined earlier in this paper. We combined the linkage (PageRank) and content scores for this experiment using both the parameter technique (linkage weight of 0.25) as well as the scarcity-abundance technique. These are referred to as 'PageRank A' (parameter combination) and 'PageRank B' (scarcity abundance) in the results section.

A more thorough description of the experiments on the WT-Dense collection is documented elsewhere (Gurrin and Smeaton 2003).

3.3.2 WT-Dense Results

Given that our linkage experiments were based on re-ranking content-only results by utilising linkage information, we can directly evaluate the benefit of incorporating linkage information into the retrieval process. In Table 2 we present a brief summary of our results, in which we see that some linkage experiments actually achieved small improvements in precision (shown as bold/italic) over content-only runs when examined at rank positions 5, 10 and 15. This is encouraging because previously, TREC participants (apart from the one experiment in 1999) were unable to obtain any improvement in retrieval performance when using WT10g data. Although WT-Dense is not the same dataset as WT10g, we are using a subset of both the dataset and the relevance judgements and when we ran many similar experiments on the full WT10g dataset (Gurrin and Smeaton 2000) we did not achieve any improvement in retrieval performance at all.

Table 2. Precision values for the experiments on WT-Dense

	Content	Citation A	Citation B	SpreadAct	PageRank A	PageRank B
5	0.2389	***0.2444***	***0.2500***	0.0500	***0.2444***	***0.2444***
10	0.1833	0.1833	***0.1861***	0.0639	0.1833	0.1806
15	0.1611	***0.1630***	***0.1630***	0.0722	***0.1630***	0.1593
20	0.1500	0.1472	0.1486	0.0750	0.1486	0.1486
30	0.1167	0.1148	0.1148	0.0787	0.1130	0.1139
100	0.0444	0.0444	0.0442	0.0406	0.0444	0.0444

Our experiments suggest that by increasing the density of off-site links within a collection it is possible to incorporate linkage analysis into retrieval and (at least) not significantly harm the retrieval process. Due to the fact that our improvements are quite small and on a very small collection using a small number of queries, our findings taken in isolation do not prove much, however when combined with the findings of some participants in TREC-2002 using the .GOV collection (discussed below) they suggest that increasing (primarily) the off-site linkage density within a test collection can help to illustrate any benefits (if any at all) that can be found by incorporating linkage-based retrieval.

3.4 The .GOV Collection (TREC-2002)

The .GOV collection used in 2002 (and 2003 also) consists of 1,247,753 documents (not all HTML) from a fresh crawl of web pages made in early 2002. It was hoped that fresher web data would better reflect today's WWW. Findings for TREC-2002 illustrate that for some participants, the application of linkage analysis did indeed improve retrieval performance in the new Topic Distillation task (Wu et al. 2002) as well as (once again) in the Named Page finding task. Recall that the Topic Distillation task was evaluated using the measure of precision at 10.

So what was the essential difference between the .GOV collection and the previous TREC collections? The off-site link density of the .GOV collection (averaging 1.98 off-site in-links for each document) was far greater than that of WT10g (0.1) and WT2g (0.01). This, we believe, was the primary reason for the encouraging findings in TREC-2002, however, as experiments we present later suggest, off-site inlink density of .GOV is only about 40% of the off-site in-link density of documents on the WWW. In addition, the nature of the search task had changed, from the conventional ad-hoc retrieval of previous web tracks to the new topic distillation task, which may also have had some influence on findings, however our focus is on linkage issues.

In summary, the web track in TREC has contributed in a major way to focusing attention on scientifically measuring the effectiveness of web

searching but progress in actually achieving that measurement has been slow, principally, we believe, because of the test collections used. A dataset better capable of supporting linkage-based retrieval should more clearly illustrate if any benefits are possible by integrating linkage analysis into the retrieval process. The .GOV collection was an important step on the way but as we will show, it does not meet criteria we lay down for the composition of a truly representative web test collection.

So, given that increasing the (off-site) linkage densities within test collections seems to influence linkage algorithms, what should the criteria be for a collection that faithfully supports evaluation of linkage-based retrieval? If we can identify criteria that aid in the generation of representative test collections that more accurately reflect the complex linkage structure of the WWW, then it would be possible to illustrate more clearly if, and by how much, linkage-based techniques can aid conventional retrieval performance. In order to identify these criteria we had to examine the linkage structure of the WWW.

4. EXAMINING REAL-WORLD WEB STRUCTURE

Before examining the structure of the WWW it is essential to have a clear understanding of what aspect of web structure is to be examined, either the actual web structure or the crawlable (visible) web structure can be examined. The visible web is that section of the WWW that a conventional web crawler can reach by following the hyperlinked structure of the WWW from a reasonable starting set of web pages. The invisible web refers to those sections of the WWW that a conventional web crawler will never find, either because they are not linked into the main body of the WWW, or because they are non-static pages that exist only on databases and are dynamically generated (the dynamic web).

4.1 Methods of Examining WWW Structure

Let us look briefly at two techniques that can be used to examine web structure, sampling the actual WWW and sampling a large crawl of the WWW.

4.1.1 Sampling the Actual WWW

If the actual WWW is examined, the findings should accurately reflect true WWW structure, for both the visible and invisible sections of the WWW. One possible technique to do this is to generate random IP addresses

and if a web site is found at that address, then to examine a number of pages at an arbitrary depth down the linkage graph of the website. In this way, all pages on the WWW that are accessible from the root page of a web site will have a chance of being sampled, even though not all these pages may be found by conventional web crawlers due to the linkage structure of the WWW.

The aim of the research presented here, however, is not to specify the requirements for a test collection that accurately synthesises the web, both visible and invisible, on a micro scale, rather we aim to specify a set of requirements that will aid the synthesis of test collections that recreate the complex linkage structure of the visible web, as crawled by search engine web crawlers. Remember that all search engines' indexes are inevitably based on a large web crawl. Therefore, given that the final goal of this research is to recreate the structure of the visible WWW within a small-scale test collection, sampling the actual WWW is not necessary, rather, it is important to examine just the visible WWW structure. Our methodology for doing this is to examine web crawls and therefore throughout the rest of this paper when we discuss sampling, it refers to sampling the visible web.

4.1.2 Sampling a large crawl of the WWW

Much of the previous work in the area of examining web structure to identify linkage densities and distributions, such as (Broder et al. 2000, Soboroff 2002), have focused on examining the output of a given set of crawled web pages (i.e. the visible web). If examining crawled data to generate linkage statistics, the crawl should ideally be as large as possible in order to generate accurate (indegree especially) linkage statistics. Ideally the crawl should gather every document (i.e. the crawl should be exhaustive) which is not feasible. Search engines such as Google and AllTheWeb both index three billion web pages[5], which although not exhaustive, will represent a large proportion of the WWW and likely the densely linked core of the WWW.

In the case of Broder et al. the crawl upon which they base their linkage structure findings was 203 million web pages in size. Therefore, when calculating indegree statistics this was still being done on a subset of the actual WWW linkage graph so this will inevitably underestimate the actual indegrees of documents on the WWW. Our experiments (detailed later) find that only 25% of links parsed from a large collection of (94.5 million)

[5] Google (GOOGLE 2003) claims to index 3,307,998,701 web pages and AllTheWeb (ALLTHEWEB 2003) claim to index 3,151,743,117 web pages, both as of 1st September 2003.

documents actually point back at documents in the collection. In effect, 75% of the possible linkage density within the collection is being lost, which strongly suggests that any indegree statistics generated by examining inlinks into crawled data underestimates the actual indegree densities unless the collection used is very large indeed. Of course, if the crawl was large (for example a Google web crawl) the linkage graph would be far more complete and the indegrees of documents would be more representative of the actual figures.

In order to avoid this problem of underestimating indegree statistics, our approach was to estimate indegree linkage densities by examining outdegree linkage densities (taking care to exclude broken links from our calculations) from both a small sample of web data and a web crawl of 94.5 million documents. In the first experiment (outlined below) we describe the results of a small-scale survey of web pages undertaken in early 2002 and in the second experiment we describe our examination of the linkage structure of a much larger collection of crawled documents from 2001.

4.2 A Small scale Survey of Web Linkage

We present the results of a small survey of web pages from early 2002, the aim of which was to explore the linkage structure of the WWW in order to estimate the average off-site and on-site indegree figures for web pages. Previous work has been carried out in the area of small-scale web sampling, an example being the SOWS III survey of web structure (SOWS 2003) and experiments carried out by Cyveillance Inc. (CYVEILLANCE 2003) in 2000. Our findings presented below are very similar to the findings SOWS III (and similar to Cyveillance with regard to off-site links more so than on-site links), for more details see (Gurrin and Smeaton 2003).

For the purposes of identifying criteria to which representative test collections should adhere, we needed to examine the valid (non-broken) off-site and on-site link structure of documents from a large web crawl.

4.2.1 Surveying the WWW

When generating our own sample we identified the target population as being all visible web pages, i.e. a very large web crawl. Our approach to generating random URLs required the use of a web accessible random URL generator (UROULLETTE 2003). Our sample size was 5,000 and if we examine our sample at a 95% confidence level, our confidence interval is 1.39%, which compares favourably with a confidence interval of 6.93% for SOWS III. This means that we can be 95% confident that our results are +/- 1.39% of our stated figures. One caveat with our survey is that these

confidence figures assume truly random sampling and since we relied on URouLette for our random URLs, we are not sure how random our sample is and from how large a list of candidate URLs the random URLs are chosen. All sampling techniques that rely on choosing a document at random from a web crawl are only random with respect to the crawled data and thus cannot be classified as truly random, but as we have stated, sampling crawled data (visible web pages) was our goal.

In order to estimate in-degree densities, all out-links (HTTP only) from the 5,000 web pages were immediately parsed and all target web pages immediately cached for future examination. Since the WWW is basically a directed graph, each out-link is also an in-link, and (as stated) by observing the average outdegree of each document it is possible to identify the average indegree of every document, assuming that broken links have been identified and removed from the calculation.

4.2.2 Observations from our Survey

Based on our survey we are in a position to present some findings on the link structure of these 5,000 documents. We find that, 3,940 documents contain HTTP-only out-links. This figure comprises 2,706 documents containing off-site outlinks and 3,571 documents containing on-site outlinks. The average (HTTP-only) outdegree of WWW documents we found to be 19.8, which is composed of 5.2 off-site links and 14.6 on-site links. However, after examining all of the downloaded target pages (to identify broken links) we found that 3.2% of all out-links were broken links in that they yield only an HTTP 404 type error or equivalent.

We can identify the following valid (non-broken) out-link structure for all documents on the web (rounded to one decimal place) in Table 3.

Table 3. Average document outdegree figures from our survey

Average off-site out-degree for each document	4.9
Average on-site out-degree for each document	14.2
Average out-degree for each document	19.1

In order to address concerns that our sample may have been too small to accurately reflect WWW linkage outdegrees, we have also examined a large crawl of web data in anticipation of finding similar results, which would validate the results of this small survey.

4.3 Examining a Large Web Crawl

We have also examined a large web crawl, the SPIRIT collection. SPIRIT is a collection of 94,552,871 documents that was crawled in 2001 by University of Waterloo, Canada.

For our examination, all 2,132,947,009 links (excluding mailto links which had previously been removed), were examined and if found to be HTTP links, were classified as being either off-site or on-site. In all, we found that HTTP links accounted for 98.7% of all links processed. See Figure 1 for a basic protocol breakdown.

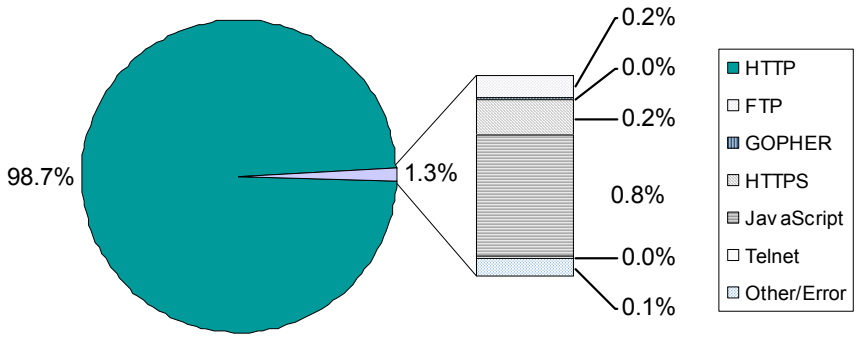

Figure 1. A basic protocol breakdown for the SPIRIT collection

Table 4 illustrates some findings about link targets from processing all HTTP links from the 94.5 million documents. 'All links' refers to all links parsed from SPIRIT, regardless of the whether the target document is contained within the collection, while 'closed links' refer to only those parsed links (499,994,497) that have both source and target URLs contained within SPIRIT. In total there were 117,644,697 closed off-site links and 382,349,800 closed on-site links. In effect, almost 75% of outlinks do not qualify as closed links.

Table 4. Examination of the proportion of on-site to off-site links within the SPIRIT collection for both all links (2,132,947,009) and closed links only (499,994,497)

	Off-site	On-site
All Links	24.73%	75.27%
Closed Links	23.53%	76.47%

A further examination of 50 million links from the collection had suggested that 92.3% of links point to recognisable HTML pages as opposed to other types of document, such as PDF files or image files. By extracting 92.3% of the HTTP-only links as being to HTML files (or equivalent) and removing 3.2 % of links that we estimate from our previous sample to be broken, we find that the average off-site out-degree figure is 4.9 and the average on-site outdegree figure is 14.7. These figures from processing the SPIRIT collection are remarkably similar to the findings of our survey of 5,000 web pages. There is a small difference in the on-site indegree, but this is not large enough to be a cause for concern. Therefore, even though our sample was only of 5,000 web pages, we are confident of the accuracy of the sample. There is, however, an additional test of accuracy, which is based on examining the distribution of outlinks from the sample, which we discuss later.

If we examine these figures for the in-links (i.e. just examine the closed links) as opposed to the out-links (all links) then these figures drop to 1.24 for the average off-site indegree and 4.04 for on-site indegree as can be seen in Table 5. This clearly illustrates the problem of calculating indegree statistics from a relatively small set of crawled data. Clearly, the majority of links from documents within the collection point at documents outside of the collection, so although the (valid) outdegree densities are as expected, in indegree densities fall well below what would be expected.

Table 5. Actual linkage densities within the SPIRIT collection

	Off-site degree	On-site degree	Total
Out-links	4.92	14.97	19.89
In-links	1.24	4.04	5.24

4.4 Comparing these Findings to the Existing Test Collections

Assuming that all out-links (with the exception of broken-links) also act as in-links then Table 6 compares the average off-site indegree figures for the TREC collections, WT-Dense and what our survey suggests to be the ideal figure.

Table 6. Comparing our sample to recent TREC collections

	WT2g	WT10g	WT-Dense	.Gov	Web Survey
Average indegree	0.011	0.101	1.425	1.98	4.916

These findings illustrate that WT2g, WT10g, WT-Dense (our subset of WT10g) and the .GOV collections seriously underestimate the off-site link

density of the WWW and since WT10g and WT2g have an even lower off-site link density than WT-Dense or .GOV, the problems with using WT10g and WT2g are even more acute. Recall that WT2g and WT10g did not support participant's experiments into linkage-based retrieval at all. In fact, the incorporation of linkage-based retrieval was seen to significantly harm retrieval performance when using these test collections.

However as stated, simply examining the average number of links within the collection is not the only measurement that we can use. The distribution of the indegrees and outdegrees will not be uniform within any collection and examining these distributions will give us a greater insight into the topology of the collection (Soboroff 2002) and will become an important aspect of our requirements for synthesising a test collection to faithfully support experiments into linkage-based retrieval of web data.

4.5 Examining the Distribution of Web Page Outdegrees

It is known that the distribution of web page indegrees and outdegrees follow closely to a power-law distribution (Broder et al. 2000) and this is considered to be a "basic web property". We are told that the "distribution of inbound links on the web as a whole is closest to a pure power-law" while "category specific distributions exhibit very large derivations from power-law scaling" (Pennock et al. 2002). This raises issues for the generation of test collections because any attempt to influence the documents comprising a collection in order to include some category specificity (perhaps to aid in query selection) will result in problems when trying to recreate the natural web link structure. However, test collections such as those used by TREC are considered to be non-category specific and thus we can be satisfied that the distributions of document indegrees should approximate a power-law distribution, which is indeed the case for WT10g and .GOV (Soboroff 2002).

4.5.1 Power Law Distributions

Power laws are used in mathematics when one wishes to relate one quantity to the power of another. A power-law implies that small occurrences are extremely common whereas large occurrences are extremely rare. If applied to web page indegrees or outdegrees this means that the vast majority of web pages have a very small number of in (or out) -links and few pages have a large number of in (or out) -links. Power law distributions are not just used to describe the indegrees of web pages (or computer science problems in general), rather they are common to both man made and naturally occurring phenomena (Mitzenmacher 2001). From computer science we see power-law distributions in web page indegrees (Broder et al.

2000), outdegrees (Faloutsos et al. 1999), in the number of pages on websites (Adamic and Humberman 2001), in Internet growth models (Mitzenmacher 2001), and in the distributions of word frequencies in language (Adamic 2003).

Let y be the number of web pages with an indegree of x. The power-law states that that the probability of a given web page having an indegree of x is proportional to x^{-a} where $a > 1$. Therefore one can calculate the number of web pages with a given indegree using the following formula[6]:

$$y = C\,x^{-a} \tag{2}$$

This illustrates that $log(y) = log(C) - a\ Log(x)$ as shown in (Adamic 2003). The constant C is a point chosen on the power law line dictating the lines position on the log-log plot. It is intuitive to choose the Y intercept of the power law line as a value for C.

A power-law distribution with exponent a is seen as a straight line with slope a on a log-log plot. This is the characteristic signature of data that follows a power-law distribution. Previous research in this area (Broder et al. 2000) has calculated exponent values (values of a) for indegrees (2.1), off-site only indegrees (2.09), outdegrees (2.72) and off-site only outdegrees (2.67) by examining a large crawl of 203 million web pages. These findings are similar to findings of a previous experiment (Kumar et al. 1999) on a similar sized collection and to another experiment on a smaller (325,729 document) collection (Barabasi and Albert 1999), which illustrates that the exponent of the power-law slope is not dependent on the size of the collection alone. The indegree and outdegree power law exponents of both WT10g and .GOV are compared to Broder's findings in Table 7.

Table 7. Comparing indegree and outdegree exponents

	Indegree	Outdegree
Broder et al.	2.1	2.72
WT10g	2.03	2.49
.GOV	1.95	2.71

It is notable that although WT10g and .GOV follow a power law distribution (Soboroff 2002), the exponents (see Table 7) do differ sometimes substantially from the findings of Broder et al. suggesting that the linkage densities within these collections are not representative of what Broder et al. found by examining their 203 million document web crawl.

[6] Assuming the indegree is greater than zero.

4.5.2 Using the Power Law to Examine the Accuracy of our WWW Survey

The power law can be used to examine the accuracy of the findings from our sample of 5,000 web pages. Were our sample to be accurate then the distributions of our outdegree calculations should have the characteristic signature of a power-law, a straight line on a log-log plot. If we examine the distribution of non-broken off-site and on-site out-links in our web sample then we can see that both approximate a power-law distribution, as can be seen in Figure 2.

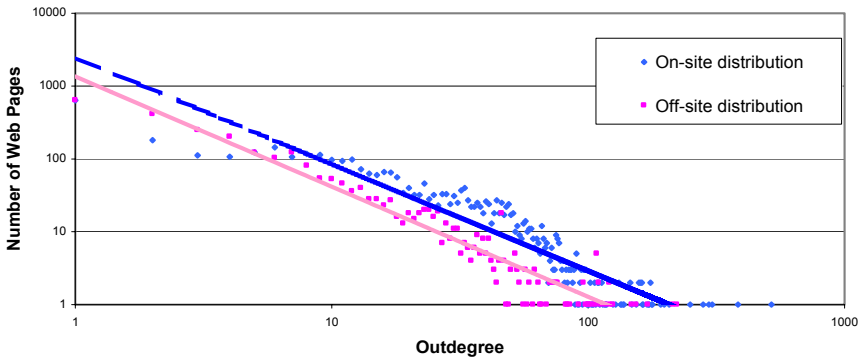

Figure 2. The off-site and on-site outdegree distributions of our sample plotted on a log-log scale with broken links removed, including trendline (correlation co-efficient = 0.9005 for the off-site distribution and 0.8542 for the on-site distributions) [7]

It is evident that both off-site and on-site outdegree distributions approximate a power-law and that this is precisely what we would expect to find if our sample was valid. An observation concerning the shape of the on-site outdegree distribution is that the tail of the on-site distribution in Figure 2 does deviate slightly from the power law. This would suggest that our sample may slightly underestimate the average on-site outdegree which is a reasonable conclusion given the average on-site outdegree figure of 14.9 for the SPIRIT collection and only 14.2 for this sample. More research is needed before we can identify if indeed our estimate of the average on-site indegree of 14.2 is (marginally) too low. However, in reality the figure of 14.2 will be

[7] The correlation co-efficient, which will range from 0..1, identifies how well the underlying linkage distributions adhere to the power law (represented by the trendline) with 1 being perfect positive correlation to a power-law distribution.

used as a guideline value for the linkage density, so this small variance should not pose any problems.

4.5.3 Problems of using the Power Law to Synthesise Test Collections

Given our understanding of power law distributions, it is valid to state that any collection, the aim of which is to faithfully support experiments into linkage based IR, should contain a certain density of links with these links being distributed according to the power law. However, a major limitation of using the power-law distribution formula $y=Cx^{-a}$ to estimate absolute linkage densities is that it only holds when $a > 0$. It is for this reason, that the probability of y documents having an indegree of x is proportional to x^{-a} and not equal to x^{-a}. As a consequence of this, the proportion of web pages that can be expected to have zero off-site and zero on-site links within the collection must be known in advance of estimating linkage densities. Once these values are known, it becomes possible to utilise $y=Cx^{-a}$ to specify absolute required linkage distributions for a document collection of size N given a certain average indegree for every document.

Further research is necessary in order to examine what proportion of web pages can be expected to have zero off-site and zero on-site links within the collection. These figures must be identified before we can specify the required distribution of off-site and on-site indegrees. It is intuitive to assume that most visible web pages will have at least one on-site in-link while most will not have any off-site in-links at all. Experiments on a small crawl of 253,922 web pages that we made in 2001 suggest that 76% of all web pages identified by parsing out-links from documents contain at least one on-site in-link, whereas only 26% of such documents contain an off-site in-link. However, the output of small web crawls is #dependent on many factors, especially the queuing algorithm employed, therefore additional experimentation is required before we specify concrete figures for documents with zero off-site and on-site indegree figures.

5. SYNTHESISING REPRESENTATIVE TEST COLLECTIONS

As a result of our examination of web structure and the notable work of the TREC web track organisers into methods of constructing a test collection (Bailey et al. 2003), we can identify the requirements for a test collection to support truly accurate investigation into linkage-based retrieval. We know that (Bailey et al. 2003) certain web properties such as size and dynamism

cannot be reflected in a small fixed test collection, however, it is essential that we carefully decide what properties to preserve or optimise.

5.1 Required Properties of a Representative Test Collection

A test collection to model real web searching should have the following properties:

- It must contain a sufficiently large and representative document set (Bailey et al. 2003).
- It must have a representative link structure within this document set.
- It must contain a large set of representative web queries (Bailey et al. 2003).
- Sufficiently complete relevance judgments (Bailey et al. 2003).
- Finally, it should be of sufficiently high generality of the dataset so as to clearly illustrate any benefit that linkage-based retrieval techniques bring to web retrieval.

The required link structure of a representative collection can be thus summarised:

- Must have an average off-site indegree adequately near to 4.9.
- Must have an average on-site indegree adequately near to 14.2.
- The indegree distributions (both off-site and on-site) should approximate a power-law distribution.

5.1.1 Power law Requirements of Representative Test Collections

Previous research, as we have seen, states that the exponent for the power law distribution of off-site indegrees is 2.09 and the exponent for the distribution of all indegrees is 2.1. A naïve approach to synthesising a representative test collection would be to simply estimate linkage density distributions using $y=Cx^{-2.09}$ for off-site indegree distributions and $y=Cx^{-2.1}$ for total indegree distributions. This will not be successful because the density of links synthesised will fall far below the required expected figures of 4.9 off-site outlinks and 14.2 on-site outlinks. Recall that the number of closed links within the SPIRIT collection was just over 25% of the total, which suggests that the indegree exponents 2.09 and 2.1 will produce a distribution containing too few inlinks. Rather, one would expect the exponents to actually be closer to the outdegree exponents (2.67 and 2.72) than the indegree exponents.

As stated, the formula $y=Cx^{-a}$ only applies to positive values of x. When we estimate that each document has an average off-site indegree of

4.9, this average includes documents with no off-site indegree at all, which we expect to be the majority of documents. While off-site indegrees will pose a major problem, the problem is far less acute with regard to on-site indegrees as virtually all documents (especially those found by a large web crawl) will have a positive on-site indegree. We plan to do additional experiments in order to generate proportions of web documents from a search engine index that have zero off-site and on-site indegrees. Once these figures are identified, the distributions can be specified more accurately.

5.2 Synthesising a Representative Test Collection

Synthesis of such a test collection is a complex task, primarily because of the complex linkage structure that needs to be modeled, but also because of the presence of heterogeneous documents, dynamisity of the content and rapid evolution and growth of the WWW itself. However, notwithstanding these problems, there are two obvious techniques could be used to synthesise such a test collection, both of which have actually been used in generating the two most recent test collections for the TREC Web track. We describe these techniques as the crawling technique and judicious selection technique.

5.2.1 Crawling a Representative Test Collection

Simply crawling a set of web documents using a web crawler[8] is an obvious choice to generate a small-scale snapshot of the WWW and was the approach taken in the generation of the .GOV collection. One of the essential features of a web crawler is the URL queue and its associated queuing algorithm. All URLs parsed from downloaded documents are enqueued (pushed onto the queue), unless they have already been visited or are already on the queue, and they can be dequeued (removed from the queue) in the order determined by the queuing algorithm. The queuing algorithm is extremely important, as it will determine the behavior of the web crawler and as a consequence, the resultant linkage densities of the crawled data.

Web crawls can be either exhaustive (all possible pages crawled) or non-exhaustive. In such a large and dynamic environment such as the WWW, the resources required to complete an exhaustive crawl are so large as to be prohibitive, and thus all crawls are essentially non-exhaustive.

[8] A web crawler is a software tool that gathers web pages from the WWW, usually for the purpose of examining the nature of, or providing content-retrieval facilities over, web pages.

5.2.2 Crawling issues for non-exhaustive Crawls

Given that a queuing algorithm will dictate in what order URLs are dequeued, the algorithm becomes irrelevant for an exhaustive crawl, but not so for a non-exhaustive crawl. In such a crawl, the results will vary greatly if different queuing algorithms are employed. This means that a queuing algorithm is especially influential when generating small (manageable) collections, such as those used in a TREC style environment and the test collection that this research envisages. Therefore, if queuing algorithms can generate crawled data with high off-site (and on-site) closed linkage densities, then crawling a representative test collection could become a viable option.

5.2.2.1 Existing Crawled Collections

When examining the linkage structure of crawled data to validate compliance with our requirements it is essential to only examine the 'closed links' and not all links. As we have seen from the examination of the SPIRIT collection, being limited to only processing closed links results in a substantial decrease in the overall number of links within a collection, to a figure that is just over 25% of the original. This is the reason that web crawls (like SPIRIT or .GOV) that will identify large numbers of out-links only produce collections with a relatively low density of closed links.

All crawls we have looked at thus far have created collections with an off-site linkage density that is lower than our specified requirements. To illustrate that it is indeed possible to crawl collections with high off-site linkage densities we have generated some small crawls of web data using queuing algorithms that heavily weighted off-site links. We now report on the findings of these small-scale WWW crawls of web data from June-July 2001. All of these web crawls adhere to the robots exclusion protocol.

5.2.3 Small-scale WWW crawling experiments

We generated two experimental crawls, the first crawl (CRAWL A) was of 253,922 HTML documents, roughly the size of the WT2g collection. When we stopped the crawler, another 3,116,690 URLs had been identified and remained on the URL queue, therefore the queuing algorithm has been highly influential in the final composition of the dataset. The queuing algorithm that was implemented for this crawl was a weighted ageing queue[9], which weighted off-site links higher than on-site links.

[9] We refer to an ageing queue as one in which documents with low scores do not remain permanently at the end of the queue.

The second experimental crawl (CRAWL B) produced a collection of 126,996 HTML documents. When stopped, 3,087,859 other documents were identified and remained on the queue. So once again, the queuing algorithm has been influential in the final construction of the dataset. The queuing algorithm that was implemented for this crawl maximised the number of websites visited by the web crawler.

The linkage densities of these crawls are summarised in the following table (Table 8), and compared to the .GOV collection and the SPIRIT collection.

Table 8. Comparing Linkage densities of our crawls to the TREC collections

	.GOV	SPIRIT	Crawl A	Crawl B
Number of documents	1,247,753	94,552,871	253,922	126,996
Average off-site indegree	1.98	1.24	10.0	12.6

Hence, it is clearly evident that test collections with higher off-site (and on-site) linkage densities can be generated by carefully choosing a queuing algorithm. However, this is still quite a difficult process as once URLs are judiciously chosen to be dequeued based on their off-site indegree value, the natural distribution of these indegrees will naturally deviate from a power-law distribution so this is not a trivial task. Crawling representative test collections is an area that is worthy of additional research.

An alternative approach to simply crawling a set of documents is the methodology used in the constriction of WT10g where we find that the text collection was synthesised, to meet certain criteria, from a larger collection (Bailey et al. 2003). So let us examine how a judicious selection process can be employed in order to generate a representative test collection.

5.3 Judicious Selection of Documents for a Representative Test Collection

The WT10g collection was generated by judiciously choosing documents to be added into the collection from a larger collection of documents (VLC) based on predefined criteria. We refer the reader to the paper describing the construction of WT10g (Bailey et al. 2003) for details, although the criteria used in the construction of WT10g were different to those presented in this paper.

We employ the SPIRIT collection as the large collection that acts as a source of both documents and links for test collection construction. In order to process SPIRIT and use it for this experiment we built a connectivity graph (and server) for all 94.5 million pages. This was a time-consuming process, which involved the processing of over 2.1 billion links, resulting in

a graph of 499,994,497 arcs and 94,552,871 nodes. Of these 499,994,497 arcs, 23.53% of them represented off-site links and 76.47% of them represented on-site links as shown in Table 4.

This linkage graph formed the basis of our experiments into the selection of a representative test collection. We now examine two approaches to synthesising a test collection from a larger superset, these are:

- *Random Selection*: the selection of documents at random as our baseline experiment.
- *Best-Fit Selection*: the selection of documents to match our criteria for the average off-site and on-site document in degrees.

5.3.1 Random Selection of Documents

Random selection is the process whereby documents are chosen at random from a superset of documents in order to construct a test collection. As our baseline experiment we selected 500,000 documents randomly to be added into the collection from the SPIRIT collection. We did this over a number of iterations (each adding in 50,000 documents) so that we could examine the effect on linkage densities as the size of the collection grows. This serves to illustrate how effective a random selection process is at generating test collections and as a baseline against which we can compare any alterative approaches. Figure 3 illustrates the number of both off-site and on-site links within the collection as it is expanded from 50,000 documents until a collection size of 500,000 documents had been reached.

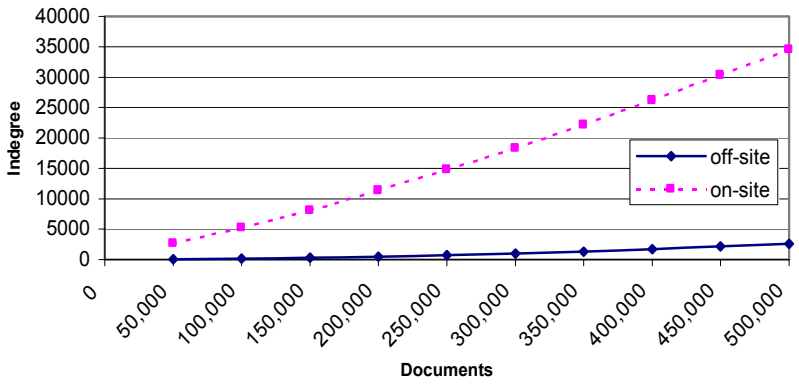

Figure 3. Graphing linkage densities as the randomly chosen collection grows in size

As would be expected, randomly selecting documents for addition into a collection will not produce a collection with the desired linkage densities. If we examine the on-site and off-site distributions (see Figure 4) we do find that they approximate a power law, albeit with different exponents to the previous research and correlation coefficients that do not suggest strict adherence to the power law distribution. We should see stricter adherence to the power law were there to be a higher density of links within the collection, or a larger number of documents in the collection.

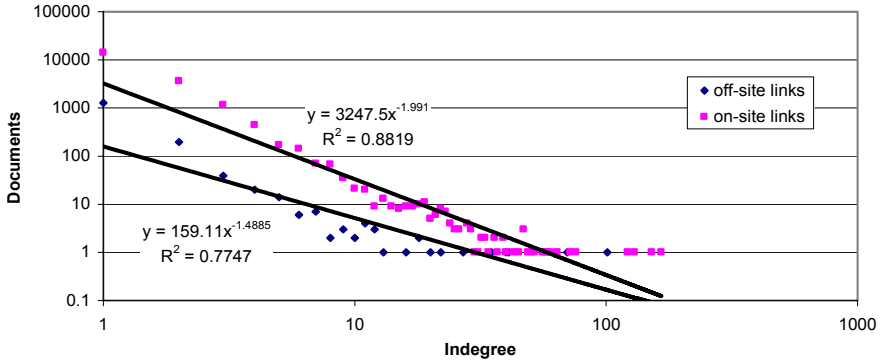

Figure 4. The Power Law distribution of the indegrees in a randomly generated collection of 500,000 documents

Although random sampling does produce a collection with a reasonable power law distribution governing the indegrees, the linkage density falls far below requirements. An alternative technique is needed, which we call best-fit selection.

5.3.2 Best-Fit Selection of Documents

Best-fit selection refers to the process of judiciously selecting documents to be added into a collection based on how these new documents will affect the current linkage graph within that collection, i.e. to meet certain predefined criteria. If we examine the methodology used in the construction of WT10g we find that the collection was generated in such a way as to fit the following criteria:

- To maximise inter-server connectivity.
- To maximise the number of pages from each server represented in the collection.
- To contain a realistic distribution of server sizes.

- And to maintain a reasonable level of generality within the collection.

We have outlined the criteria that we feel important to generate test collections earlier and as we have shown, selecting documents at random will produce a collection with an extremely low linkage density, hence some more judicious method of selecting documents is required.

Although at first, generating a test collection to meet the requirements we propose may seem relatively trivial, however we find that many unforeseen complexities arise. If the average off-site and on-site indegree figures are used as the only requirements, the distribution of indegrees will not be guaranteed to follow a power law distribution for the same reasons as the web crawls, the distribution will tend towards the average figures. How to avoid this problem is to identify how many documents with any given indegree should be contained within the synthesised collection. The formula $y=Cx^{-a}$ can be used if we can identify values for C and a, for both off-site and on-site links assuming we know the proportion of documents with zero off-site and on-site links. Once we know this, it is a fairly trivial task to synthesise the associated test collection.

In a manner similar to the construction of WT10g, we recommend dividing indegrees into bins of exponentially increasing size and identifying the probability of each document belonging to one of these bins. Then, it becomes possible, over a number of iterations, to synthesise a collection by choosing candidate documents for inclusion into the collection based on how these candidates affect the required bin distributions. The construction methodology of WT10g uses a process like this to synthesise the collection to that it meets server size distribution requirements (Bailey et al. 2003). To illustrate the process we envisage, we now describe the generation of the required distribution for a WT2g sized collection that matches our required total indegree figure[10].

5.3.2.1 Illustrating the Process for General Indegree Distributions.

As we have seen, the required average document indegree figure is 19.1. The size of the synthesised test collection we have chosen for these experiments is 262,144 (2^{18}) documents. The required distribution that we present (see Figure 5) requires the following assumption to hold:

> That all documents found by a web crawler have an indegree of at least 1.

[10] Notwithstanding the fact that we have yet to identify conclusively what proportions of documents have zero on-site and off-site indegrees.

Although this assumption will not hold for off-site indegrees, it will hold in virtually all cases for general indegrees and in the majority of cases for on-site indegrees. This is because, in order to a web document to be found by a web crawler it either has to be in the start-set (which we assume is reasonably small), or extracted from the links within a document. An examination of our own web crawls (described earlier) both show that over 99.99% of web pages in both collections had a indegree of at least one.

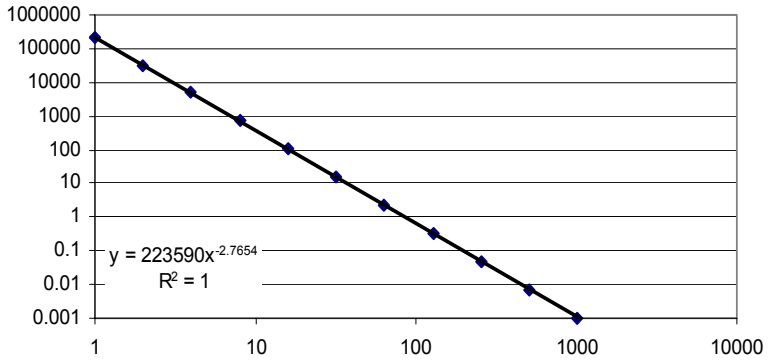

Figure 5. Graphing the power law for the indegree distribution of a collection of 262,100 documents with an average indegree of 19.1, C = 223,590 and a = 2.7654

Figure 5 illustrates a synthesised power law for the indegree distribution of a collection of 262,144 documents that have an average indegree of 19.1. From this we can generate a formula for the estimating the number of documents with a particular indegree. Let y be the number of web pages with an indegree of x and $a > 0$. Therefore can calculate the number of web pages with a given indegree (x) using formula 3:

$$y = 223590x^{-2.7654} \quad (3)$$

The number of documents that would be required to fill for each of the exponentially increasing in size bins is shown in Table 9.

Table 9. The required number of documents to fill each bin

Bin size	1	2^1	2^1-2^2	2^2-2^3	2^3-2^4	2^4-2^5	2^5-2^6
Documents	216,942	31,906	9,385	2,760	812	238	70
Bin size	2^6-2^7	2^7-2^8	2^8-2^9	2^9-2^{10}	2^{10}-2^{11}	2^{11}-2^{12}	2^{12}-2^{13}
Documents	21	6	2	1	0.15	0.04	0.01
Bin size	2^{13}-2^{14}	2^{14}-2^{15}	2^{15}-2^{16}	2^{16}-2^{17}	2^{17}-2^{18}		
Documents	<0.01	<0.01	<0.01	<0.01	<0.01		

However, for these figures to be accurate, our assumption of no documents with zero indegree must hold. Consequently more work is needed before we can actually synthesise a test collection that faithfully adheres to our requirements as outlined in a previous section. We are continuing this work with the aim of synthesising a test collection of 1.5-2 million documents, with valid off-site and on-site indegree distributions. Once this collection has been synthesised, we will reapply linkage analysis algorithms, such as the ones we applied on WT-Dense in order to answer the question of how beneficial linkage analysis techniques are for web-based retrieval tasks.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented the results of experiments into linkage-based web retrieval using both the TREC test collections and our own densely linked subset of a TREC test collection. As a result of our disappointing findings and those of other TREC participants we examined the structure of the visible web in order to identify the required linkage properties of test collection that more accurately reflects web structure than the current generation of test collection. Our belief is that only a collection that accurately recreates the linkage structure of the visible web will be able to truly answer the question of whether or not incorporating linkage evidence aids retrieval performance for web search engines. We present the findings of our examination of web structure as a set of criteria and present a methodology for the construction of a representative test collection that meets these criteria.

Although the TREC web track has aided the research field immensely by providing a framework for experimentation and retrieval performance comparison across sites, unfortunately the collections used have not supported TREC participants' experiments into linkage-based retrieval sufficiently. We have seen that although the linkage distributions of the collections follow a power law, the average inlink densities fall far below our specified requirements. However, that said, the recent .GOV collection is a major improvement, but it still falls short of the required off-site link density, in that it only contains 40% of our required off-site linkage density (see Table 6). Our synthesised WT-Dense collection also underestimates the off-site link density by a factor of over three. One question this poses is whether a certain (critical) density of off-site links is required within a collection before it can support linkage experiments. Results from the 2002 TREC web track suggest that an average of 1.98 off-site in-links into each document is sufficient to begin showing improvements in retrieval performance, but, as we have seen, even this falls well short of the ideal.

In addition to outlining the requirements for such a collection, we have identified some techniques to building the collection. Our continuing work in this area involves synthesising a representative test collection and upon construction of the collection, examining how effective this collection would be in evaluating linkage-based algorithms.

There are a number of issues related to this research that have yet to be addressed, they are:

- *Dynamic data*, we have not taken the dynamic web into account, however, much of the web's data is dynamic (automatically generated from back-end databases). Conventional search engines have to handle dynamic data, yet web test collections have thus far ignored its existence.
- *Problems of collection synthesis*, are there notable effects on a test collection by judiciously adding documents to meet a required link distribution and density? Will the documents chosen, only be the documents that are well linked into the real WWW, given that our large collection (document source) will only contain a fraction of the available web links? In addition, we have heretofore ignored outdegree densities within the collection. They will be higher than indegree densities within the collection. Does this affect the representativeness of the collection, or any linkage algorithms that utilise outlink information to aid retrieval.
- *The nature of the retrieval task*, are we evaluating linkage algorithms in the correct way? The Topic Distillation task of recent web tracks evaluates top-10 results, which is a change from the conventional ad-hoc retrieval evaluation. We feel that query analysis and consequent judicious utilisation of linkage data to suit different query types may be a key aspect of successfully incorporating linkage analysis into web retrieval. Three different types of queries, navigational, informational and transactional have been identified (Broder 2002) and each may benefit from different retrieval approaches.
- *Other requirements*, we have focused on linkage density requirements in our research, but are there other requirements that should be modeled within a test collection? For example, WT10g models server size distributions. We have, thus far, focused our efforts on linkage densities and distributions.

ACKNOWLEDGEMENTS

Access, use, and analysis of the SPIRIT Web collection was supported, in part, by the EU 5th Framework RTD project SPIRIT (Spatially-Aware Information Retrieval on the Internet: IST-2001-35047). Thanks goes to Mark Sanderson & Hideo Joho at the University of Sheffield for aiding our access to the collection.

REFERENCES

Adamic L (2003) Zipf, Power-laws, and Pareto - a ranking tutorial. http://www.hpl.hp.com/shl/papers/ranking/ (visited 1st September 2003).

Adamic L and Humberman B (2001) The Web's Hidden Order. In: Communications of the ACM, Vol. 44, No. 9, pp.55-59.

ALLTHEWEB (2003) http://www.alltheweb.com (visited 1st September 2003).

Amento B, Terveen L and Hill W (2000) Does 'Authority' mean quality? Predicting Expert Quality Ratings of Web Document. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in IR, pp. 296-303.

Bailey P, Craswell N and Hawking D (2003) Engineering a multi-purpose test collection for Web Retrieval Experiments. In: Journal of Information Processing and Management, pp. 853-871.

Barabasi A and Albert R (1999) Emergence of scaling in random networks. Science, 286, pp. 509-512.

Bharat K and Henzinger M (1998) Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in IR, pp. 104-111.

Brin S and Page L (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Proceedings of the 7th International WWW Conference, pp. 107-117.

Broder A (2002) A Taxonomy of Web Search. In: ACM SIGIR Forum, Vol. 36, No. 2, Fall 2002, pp. 3-10.

Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A and Weiner J (2000) Graph Structure in the Web. In: Proceedings of the 9th International WWW Conference, pp. 309-320.

CYVEILLANCE (2003). http://www.cyveillance.com. (visited 13th May 2003).

Faloutsos M, Faloutsos P, Faloutsos C (1999) On Power-Law Relationships of the Internet Topology. In: Proceedings of the annual ACM SIGCOMM Conference on Research and Development in Data Communications 99, pp. 251-262.

GOOGLE (2003) http://www.google.com (visited 1st September 2003).

Gurrin C & Smeaton A F. (2003) Improving the Evaluation of Web Search Systems. Advances in Information Retrieval. In: Proceedings of the 25th BCS-IRSG European Colloquium on IR Research, Springer Lecture Notes in Computer Science, pp. 25-40.

Gurrin C and Smeaton A F. (1999) Connectivity Analysis Approaches to Increasing Precision in Retrieval from Hyperlinked Documents. In: Proceedings of the 8th Annual TREC Conference, pp. 357-366.

Gurrin C and Smeaton A F. (2000) Dublin City University Experiments in Connectivity Analysis for TREC-9. In: Proceedings of the 9th Annual TREC Conference, pp. 179-188.

Hawking D (2000) Overview of the TREC-9 Web Track. In: Proceedings of the 9th Annual TREC Conference, pp. 87-102.

Hawking D, Voorhees E, Craswell N and Bailey P (1999) Overview of the TREC-8 Web Track. In: Proceedings of the 8th Annual TREC Conference, pp. 131-150.

Kleinberg J (1999) Authoritative Sources in a Hyperlinked Environment. In: journal of the ACM, Vol. 46, No. 5, pp. 604-623.

Kumar R, Raghavan P, Rajagopalan S and Tomkins A (1999) Trawling the web for emerging cyber-communities. In: Proceedings of the 8th International World Wide Web Conference, pp. 403-415.

McBryan O (1994) GENVL and WWWW: Tools for taming the Web. In: Proceedings of the 1st International WWW Conference, pp. 58-67.

Mitzenmacher M (2001) A Brief History of Generative Models for Power Law and Lognormal Distributions. In Proceedings of the 39th Annual Allerton Conference on Communication, Control, and Computing, pp. 182-191.

Murray B and Moore (2003) A Sizing the Internet - A White Paper. Cyveillance, Inc., 2000. http://www.cyveillance.com/web/corporate/white_papers.htm (visited 1st September 2003).

Page L, Brin S, Motwani R and Winograd T (1997) The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Libraries working paper, 0072.

Pennock D, Flake G, Lawrence S, Glover E and Giles L (2002) Winners don't take all: Characterising the competition for links on the web. National Academy of Sciences, Vol. 99, No.8, pp. 5207-5211.

Silverstein C, Henzinger M, Marais J, and Moricz M (1998) Analysis of a very large AltaVista query log. Digital SRC Technical Note 1998-014.

Singhal A. and Kaszkiel M (2000) AT&T at TREC-9. In: Proceedings of the 9th Annual TREC Conference, pp. 103-105.

Soboroff I (2002) Does WT10g look like the Web? In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in IR. pp. 423-424.

SOWS III: The Third State of the Web Survey (1999) http://www.pantos.org/atw/35654-a.html (visited 1st September 2003).

URouLette Random Web Page Generator (2003) http://www.uroulette.com (visited 1st September 2003).

Wu L, Huang X, Niu J, Xia Y, Feng Z and Zhou Y (2002) FDU at TREC 2002: Filtering, Q&A, Web and Video Tasks. In: Proceedings of the 11th Annual TREC Conference, pp. 232-247.